

*A COMPARISON OF MOMENTARY TIME SAMPLING AND  
PARTIAL-INTERVAL RECORDING FOR EVALUATING  
FUNCTIONAL RELATIONS*

MAEVE G. MEANY-DABOUL, EILEEN M. ROSCOE,  
JASON C. BOURRET, AND WILLIAM H. AHEARN

NEW ENGLAND CENTER FOR CHILDREN  
NORTHEASTERN UNIVERSITY

In the current study, momentary time sampling (MTS) and partial-interval recording (PIR) were compared to continuous-duration recording of stereotypy and to the frequency of self-injury during a treatment analysis to determine whether the recording method affected data interpretation. Five previously conducted treatment analysis data sets were analyzed by creating separate graphic displays for each measurement method (duration or frequency, MTS, and PIR). An expert panel interview and structured criterion visual inspection were used to evaluate treatment effects across measurement methods. Results showed that treatment analysis interpretations based on both discontinuous recording methods often matched those based on frequency or duration recording; however, interpretations based on MTS were slightly more likely to match those based on duration and those based on PIR were slightly more likely to match those based on frequency.

DESCRIPTORS: measurement, momentary time sampling, partial-interval recording

Continuous recording methods (e.g., duration and frequency) provide direct measures of response dimensions. However, they are often impractical in that they require a dedicated observer. In most clinical programs, therapists are responsible for collecting data on more than 1 individual or target response simultaneously, making continuous recording impractical. For this reason, time sampling has often been used to estimate behavior. Two often-used methods of time sampling are partial-interval recording (PIR) and momentary time sampling (MTS). PIR involves recording an occurrence if the target response occurs at any point during an interval. MTS involves recording an occurrence if the target response occurs during a prespeci-

fied moment (usually 1 or 2 s at the end of an observation interval).

The accuracy of MTS and PIR has been investigated in several studies using computer-generated behavior. For example, Powell, Martindale, and Kulp (1975) compared MTS and PIR to continuous duration for the in-seat behavior of a secretary during 20-min sessions. Results showed that PIR consistently overestimated duration, whereas MTS either over- or underestimated duration; however, the margin of error associated with MTS was much smaller than that associated with PIR. Harrop and Daniels (1986) extended the Powell et al. study by evaluating the accuracy of MTS and PIR in estimating both absolute behavioral levels and relative change in behavioral levels. The authors used computer-simulated behavior at four constant durations (1 s, 5 s, 10 s, and 20 s) and two frequency settings of low-to-medium rate and medium-to-high rate. Results showed that MTS was more accurate than PIR when estimating absolute duration. However, neither MTS nor PIR provided an accurate

---

We thank Brian Iwata for his feedback and suggestions on this project. This investigation was conducted in partial fulfillment of requirements for a Master's degree by the first author.

Address correspondence to Eileen Roscoe, New England Center, 33 Turnpike Road, Southboro, Massachusetts 01772 (e-mail: [eroscoe@necc.org](mailto:eroscoe@necc.org)).

doi: 10.1901/jaba.2007.40-501

estimate of frequency. The authors noted that PIR was more sensitive than MTS for detecting changes in the level of duration and frequency.

Suen, Ary, and Covalt (1991) discussed how the type of error obtained by MTS or PIR is a direct result of how the two sampling methods are conducted. They noted that error obtained by both methods is a result of the proportion of mixed intervals (i.e., those in which behavior occurs during only a portion of the interval). Because PIR detects all mixed intervals as an occurrence, PIR overestimates absolute duration and rate, produces biased estimates of relative change, and underestimates the magnitude of change with high-rate behavior. By contrast, MTS produces unbiased estimates of these behavioral parameters. Although these characteristics of MTS and PIR are directly caused by the method of sampling, it is still important to determine the degree to which these characteristics of recording methods influence clinically relevant data.

Most studies that have evaluated the accuracy of time sampling have used simulated behavior. A notable exception was the study conducted by Murphy and Goodall (1980), who compared MTS and PIR to continuous duration using videotaped samples of children with mental retardation who exhibited stereotypy. Results showed that MTS produced more accurate estimates of duration than did PIR. Gardenier, MacDonald, and Green (2004) extended research in this area by comparing MTS and PIR for estimating continuous duration of stereotypy among children with pervasive developmental disorder (not otherwise specified) or autism. They also found that PIR consistently overestimated the duration of stereotypy, whereas MTS sometimes overestimated and at other times underestimated duration. PIR was found to produce much larger deviation from duration recording than MTS. In addition, MTS was found to produce more accurate estimates across low, moderate, and high levels of stereotypy.

Although Murphy and Goodall (1980) and Gardenier *et al.* (2004) extended research in this area by showing that MTS resulted in less measurement error than PIR for estimating duration of a clinically important behavior, they did not compare MTS and PIR to frequency recording. Because clinicians often use discontinuous recording methods, such as PIR, for discrete responses that are appropriately measured using frequency, it may be helpful to evaluate whether MTS and PIR produce accurate estimates of frequency for clinically relevant behavior. In addition, because previous research has evaluated MTS and PIR during a baseline condition only, it is unclear whether the obtained differences in measurement error may lead to different decisions regarding treatment. For example, when evaluating treatment for an individual's stereotypy or self-injury, it may be more conservative to use a method that consistently overestimates behavior rather than to use one that sometimes overestimates and sometimes underestimates behavior, even if the former results in a greater margin of error. Thus, it is unclear whether the overestimation associated with PIR would affect the evaluation of treatment success. For this reason, it is relevant to evaluate to what extent different measurement methods may result in different treatment interpretations when evaluating trends in data paths across conditions of a treatment analysis.

In the current study, we replicated and extended previous research by comparing MTS and PIR to duration records of stereotypy and to frequency records of self-injurious behavior across treatment analysis conditions to determine whether the recording methods might affect interpretations regarding functional control.

## METHOD

### *Participants and Setting*

Four individuals who had been diagnosed with autism, and who had been referred for the

assessment and treatment of their problem behavior, participated. Amy was a 7-year-old girl who engaged in vocal stereotypy that consisted of noises and some word approximations. Daniel was a 21-year-old man who engaged in vocal stereotypy that primarily consisted of noncontextual words and phrases. Beth was a 14-year-old girl who exhibited motor stereotypy that consisted of rocking, hand flapping, jumping, and posturing. Jack was an 8-year-old boy who exhibited self-injurious behavior in the form of head banging and self-biting. All sessions were conducted in a room (1.5 m by 3 m) equipped with a videocamera, a table, and chairs.

#### *Response Measurement and Interobserver Agreement*

Vocal stereotypy was defined as any instance of noncontextual or nonfunctional speech and included babbling, singing, and phrases unrelated to the stimulus context (e.g., repeating the word "red" in the absence of a red stimulus). Motor stereotypy was defined as any form of nonfunctional movement. Examples included hand flapping, squeezing eyes shut, posturing, pressing hands onto body, jumping, or shaking head. Observers recorded vocal and motor stereotypy using the continuous duration method. Episodes of vocal and motor stereotypy were recorded during the first second a response was observed and ended when there was 1 s free of responding. Observers used a counter on the videocassette recorder that displayed the number of seconds that had elapsed from the start of the session. The occurrence of stereotypy was scored during 1-s bins on a data sheet containing 300 1-s bins (for 5-min sessions). The total number of seconds of stereotypy in each session was divided by the total number of seconds in the session and multiplied by 100% to calculate the percentage of the session in which stereotypy occurred.

Self-injury (Jack only) included self-biting, head-to-object hitting, and hand-to-head hit-

ting. Self-biting was defined as any instance of one's teeth closing around any part of the hand. Head-to-object hitting was defined as any instance of forceful contact between one's head and a stationary object (e.g., floor, wall, desk). Hand-to-head hitting was defined as striking one's face or head with an open hand or closed fist with a distance greater than 6 in. Observers scored Jack's self-injury using frequency recording. To allow PIR and MTS to be derived from the data record, frequency data were also recorded within 1-s bins using the same data sheet used for duration recording. Sessions for Amy, Daniel, and Beth lasted 5 min; sessions for Jack lasted 10 min.

Interobserver agreement was calculated by having two observers independently score sessions from videotape. For participants who exhibited stereotypy (Amy, Daniel, and Beth), point-by-point agreement was calculated by having observers score the occurrence or non-occurrence of stereotypy within 1-s bins. The number of intervals with an agreement was then divided by the number of intervals with an agreement plus the number of intervals with a disagreement and multiplied by 100%. For Jack, who exhibited self-injury, proportional agreement was collected by dividing the smaller number of responses by the larger number within each 1-s bin; these fractions were then averaged across the session and multiplied by 100%. Agreement was measured across all conditions during 33%, 67%, 31%, and 33% of sessions, for Amy, Daniel, Beth, and Jack, respectively. Agreement averaged 95% (range, 93% to 97%) for Amy, 96% (range, 78% to 100%) for Daniel, 91% (range, 68% to 98%) for Beth, and 99% (range, 96% to 100%) for Jack. In addition, occurrence and nonoccurrence agreement data were calculated by including only those intervals in which the primary observer scored an occurrence or a nonoccurrence, respectively. Occurrence agreement averaged 97% (range, 93% to 100%) for Amy, 80% (range, 60% to 100%)

for Daniel, 88% (range, 68% to 100%) for Beth, and 83% (range, 38% to 100%) for Jack. Nonoccurrence agreement averaged 81% (range, 64% to 89%) for Amy, 98% (range, 83% to 100%) for Daniel, 73% (range, 67% to 100%) for Beth, and 99% (range, 98% to 100%) for Jack.

### *Treatment Analysis*

Measurement methods were compared by using data from five previously conducted treatment analyses that had been videotaped. Data for Amy have been previously published (see Alice in Ahearn, Clark, MacDonald, & Chung, 2007). An ABAB design was used to evaluate experimental control for each treatment analysis. Three data sets consisted of an evaluation of response interruption and redirection, and two data sets consisted of an evaluation of noncontingent reinforcement (NCR) or noncontingent escape (NCE). The latter two data sets were conducted concurrently in a multielement design with Jack. The data were extracted from a combined multielement reversal design to create two separate ABAB data sets for use in this experiment.

*Response interruption and redirection.* Amy, Daniel, and Beth received this intervention because they exhibited problem behavior (stereotypy) maintained by automatic reinforcement (based on the results of previous functional analyses). During the baseline (A) phase of the reversal design, no-interaction sessions, in which a therapist was present in the room and offered no materials or interaction, were conducted. There were no programmed consequences for stereotypy during this condition. During the treatment (B) phase of the reversal design, response interruption was used, in which instances of stereotypy immediately resulted in contingent instructions to complete brief vocal or motor tasks. For example, if the participant engaged in vocal stereotypy, the therapist prompted eye contact and then issued prompts to engage in an appropriate vocal response (e.g., "What color is the table?"). If the

student engaged in motor stereotypy, the therapist prompted eye contact and then issued prompts to engage in an appropriate motor task (e.g., "touch your toes" or "touch your head"). Original treatment decisions regarding when to change phases were based on duration recording.

*NCR or NCE.* Jack received this intervention because he exhibited problem behavior (self-injury) maintained by escape from task demands (based on the results of a previous functional analysis). During the baseline (A) phase of the reversal design, NCR without extinction and NCE without extinction were evaluated. During both NCR and NCE, demands were continuously delivered and extinction was not in effect (i.e., self-injury continued to result in a 15-s break). In addition, during NCR, preferred edible items were delivered on a fixed-time (FT) 15-s schedule; during NCE, a 15-s break was delivered on an FT 15-s schedule. During the treatment (B) phase of the reversal design, an extinction component (i.e., self-injury no longer resulted in escape) was added to the NCR and NCE interventions. Original treatment decisions regarding when to change phases were based on frequency recording.

### *Data Analysis*

To evaluate whether the different measurement methods resulted in similar interpretations, separate graphic displays were created for each measurement method. Duration or frequency recording data were used as the standard of comparison, and MTS and PIR were derived from the original data record. To obtain PIR data, the original data sheet was segmented into 30 10-s intervals. The observer noted whether a response was recorded within each 10-s interval. If any responding was recorded, an occurrence was scored. If no responding was recorded during the interval, a nonoccurrence was scored. To obtain MTS data, the observer scored an occurrence if responding was recorded during the 2 s following an interval (e.g., an occurrence was recorded if responding was

observed during Seconds 11 and 12, Seconds 21 and 22, Seconds 31 and 32). If responding was not observed during either of the specified seconds, a nonoccurrence was scored.

As noted previously, duration data were summarized as percentage duration by dividing the number of seconds in which a response was recorded by 300 (the total number of seconds in a session) and multiplying by 100%. Frequency data were summarized as responses per minute by dividing the total number of occurrences by the number of minutes in a session. PIR and MTS measures were summarized as percentage of intervals by dividing the number of intervals in which an occurrence was scored by the total number of intervals and multiplying by 100%.

For each of the five data sets, three graphs were created (one displayed the data when measured using either duration or frequency, one displayed the data when measured using MTS, and one displayed the data when measured using PIR), resulting in a total of 15 graphs. Figures 1, 2, and 3 show the results from Amy's, Daniel's, and Beth's treatment assessments, respectively. Figures 4 and 5 show results from Jack's NCR and NCE treatment assessments, respectively. The range of the scale for the  $y$  axis was standardized across graphs by identifying the highest data point across all conditions, rounding that value up to the nearest multiple of five, and using this number for the maximum  $y$ -axis value. After the graphic displays were created, two methods were used to evaluate treatment effects: an interview and a structured criterion visual inspection as described by Fisher, Kelley, and Lomas (2003).

*Expert panel interview.* Nine individuals served as an expert panel. All panel members had a minimum of a master's degree in behavior analysis, were currently board-certified behavior analysts, served as faculty in an applied behavior analysis graduate program, had a minimum of 5 years of clinical experience, and had extensive experience making treatment decisions based on visual inspection of data.

During the expert interview, the first author met with each of the informants individually. She presented each of the 15 graphs consecutively to the panel member and asked him or her to inspect the graph. While he or she was viewing each graph, the panel member was asked, "Is there an overall demonstration of functional control?" and was instructed to respond "yes" or "no." No other instructions or information was given to the panel members. Informants' responses were recorded on data sheets by the first author.

Data obtained from this interview were analyzed by comparing informants' responses for the MTS or PIR data sets to their responses for the continuous recording (duration or frequency) data sets. If the same response was obtained for both data sets, an agreement was scored.

*Structured criterion method.* The structured criterion method was based on the dual criterion method developed by Fisher et al. (2003). Quantitative parameters associated with each data set were used to create a series of lines for each graph to aid in determination of functional relations. First, data were divided into two equal parts per baseline phase, and lines were drawn vertically and horizontally at the midpoint of each section. Second, another line was drawn connecting the midpoints of each section to create the quarter-intersect line of progress. The quarter-intersect line of progress was moved up or down so that the distribution of data points on either side of the line was equal. This line was then referred to as the split-middle line of progress. To evaluate changes across phases (i.e., from the baseline phase to the treatment phase), the split-middle line of progress and the mean line based on each baseline condition were drawn on each of the subsequent treatment phases. A treatment effect was recorded if a prespecified minimum number of the data points, based on the criteria proposed by Fisher et al., in the treatment phase fell away from the split-middle line and from

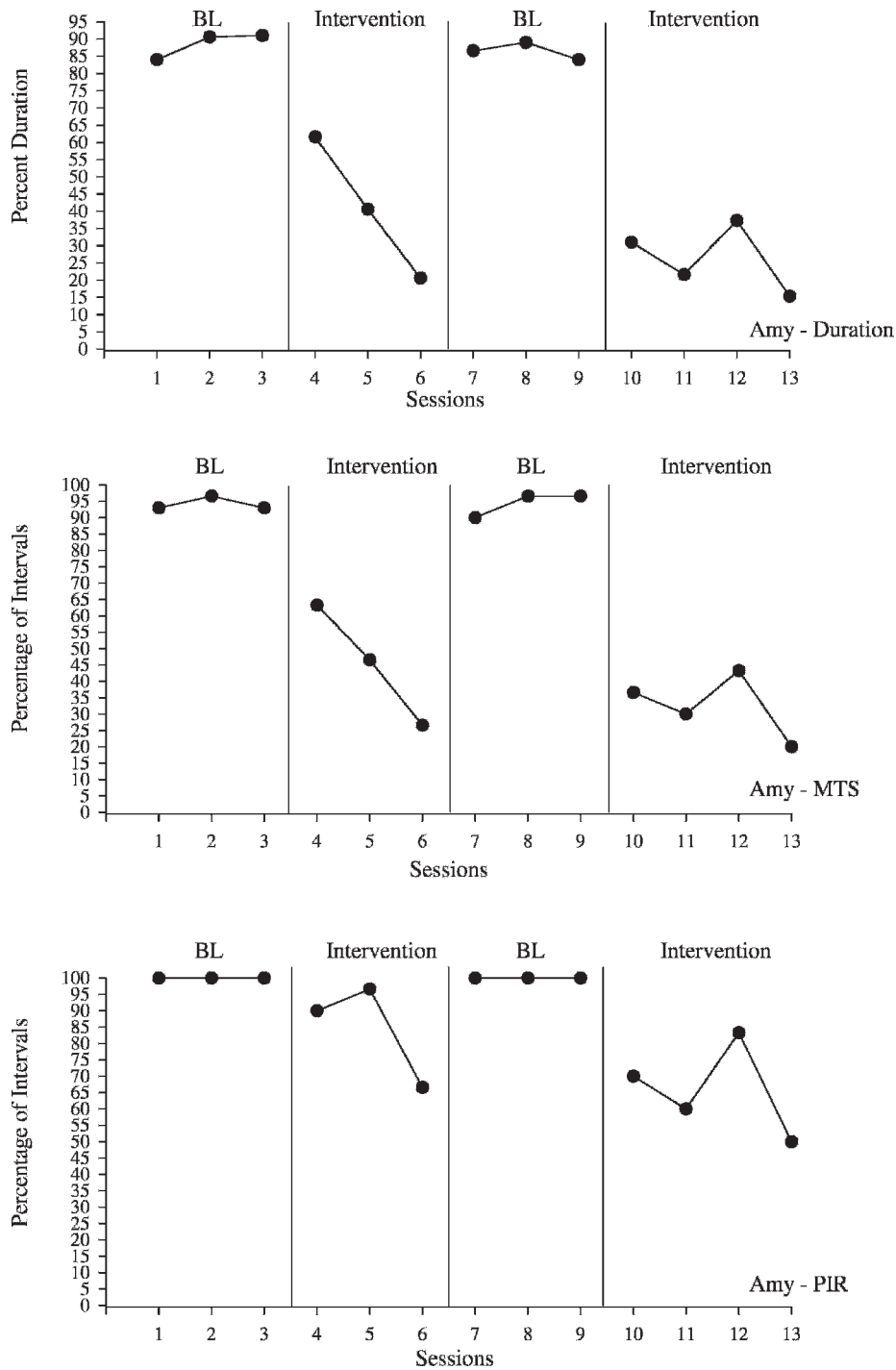


Figure 1. Intervention data (response interruption and redirection) based on duration, MTS, and PIR for Amy.



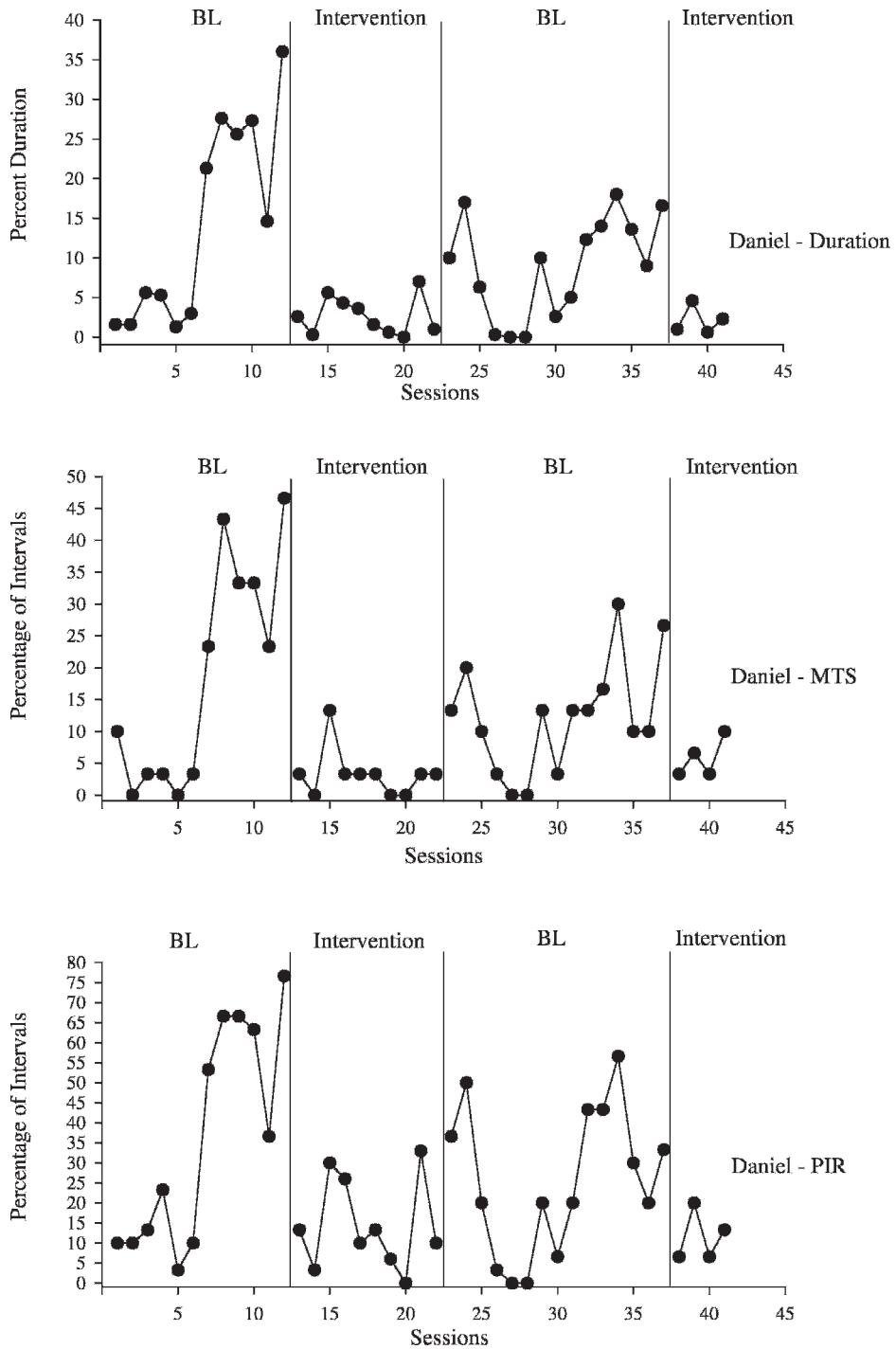


Figure 2. Intervention data (response interruption and redirection) based on duration, MTS, and PIR for Daniel.

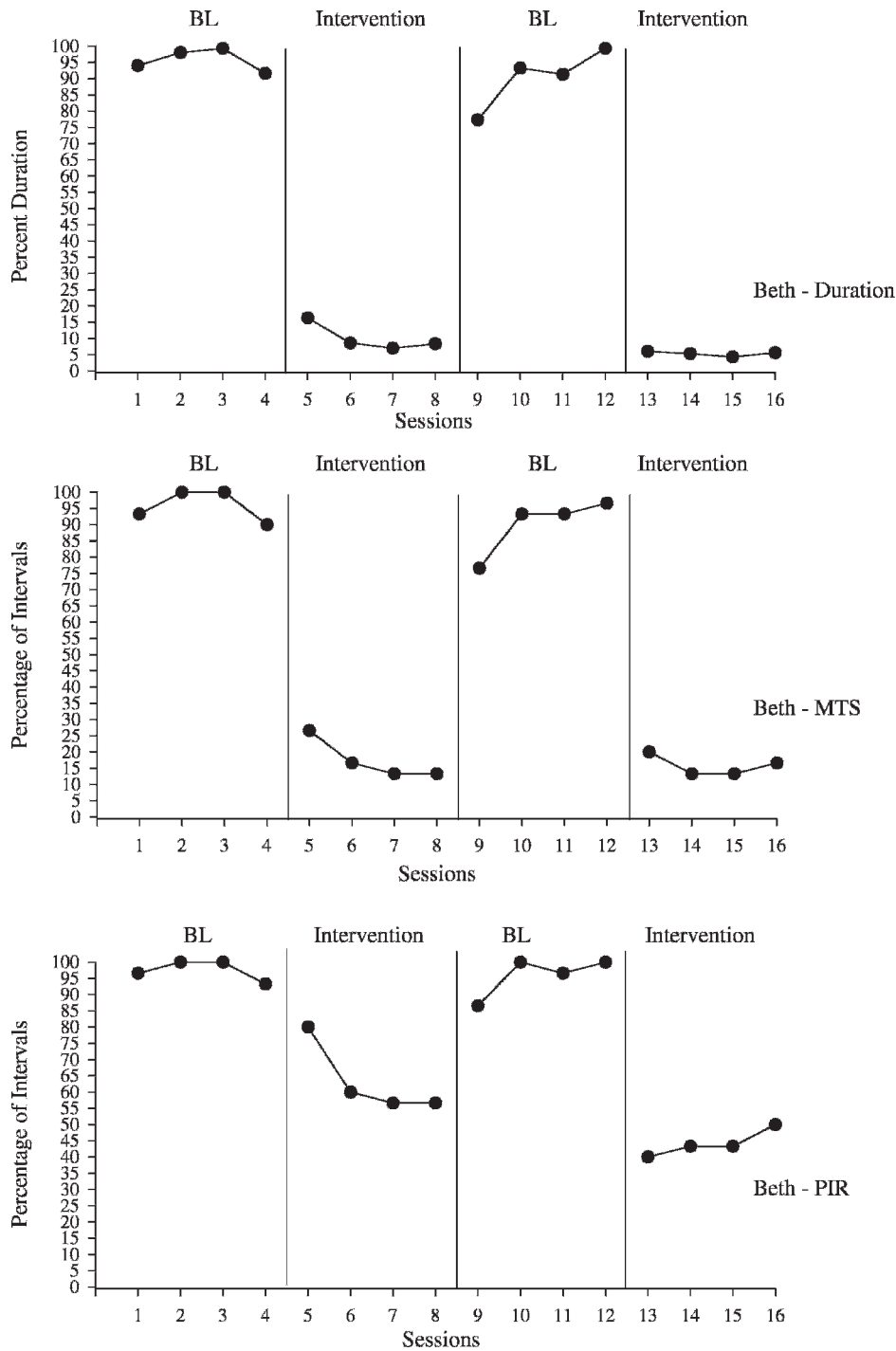


Figure 3. Intervention data (response interruption and redirection) based on duration, MTS, and PIR for Beth.



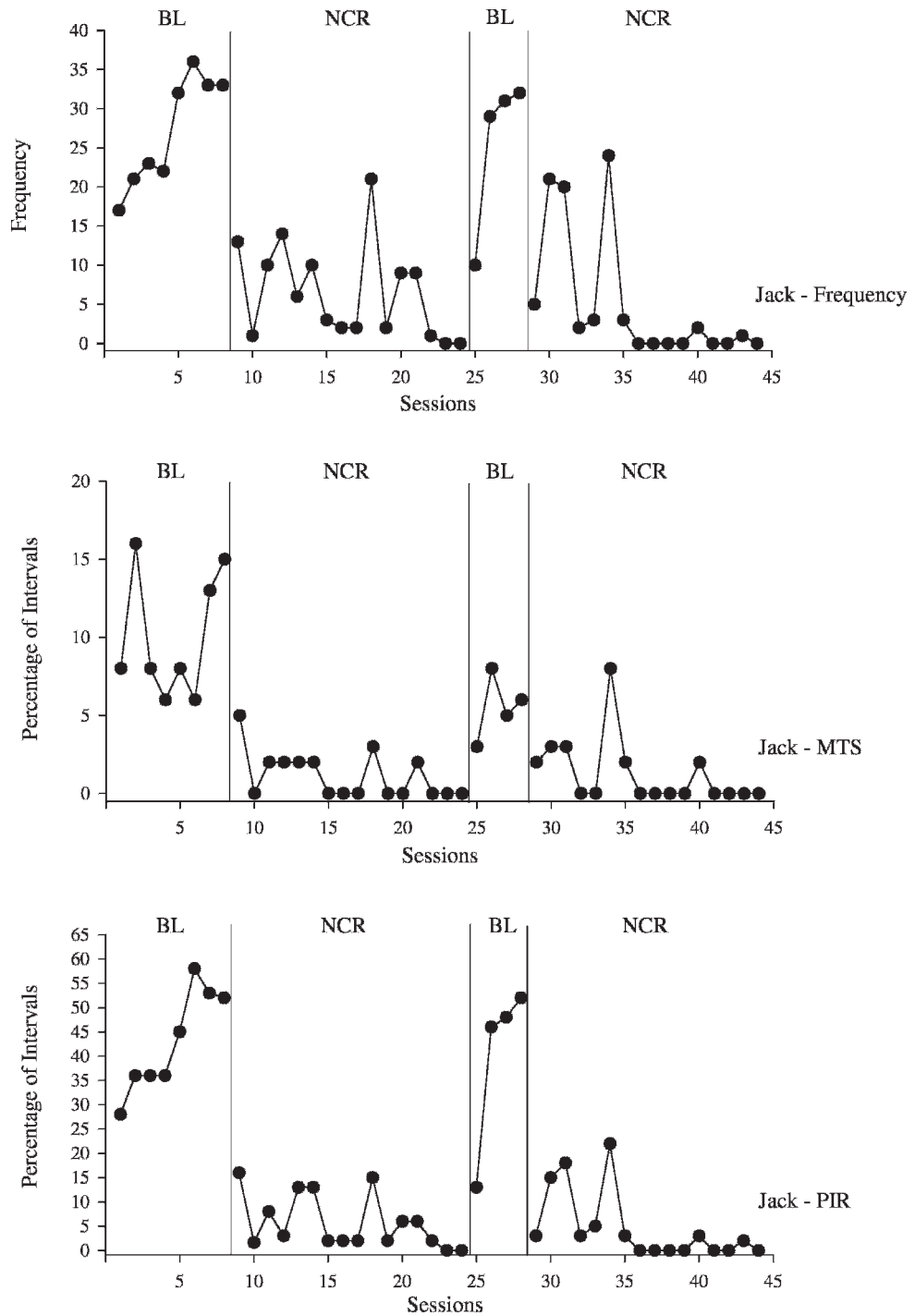


Figure 4. Intervention data (NCR with extinction) based on frequency, MTS, and PIR for Jack.

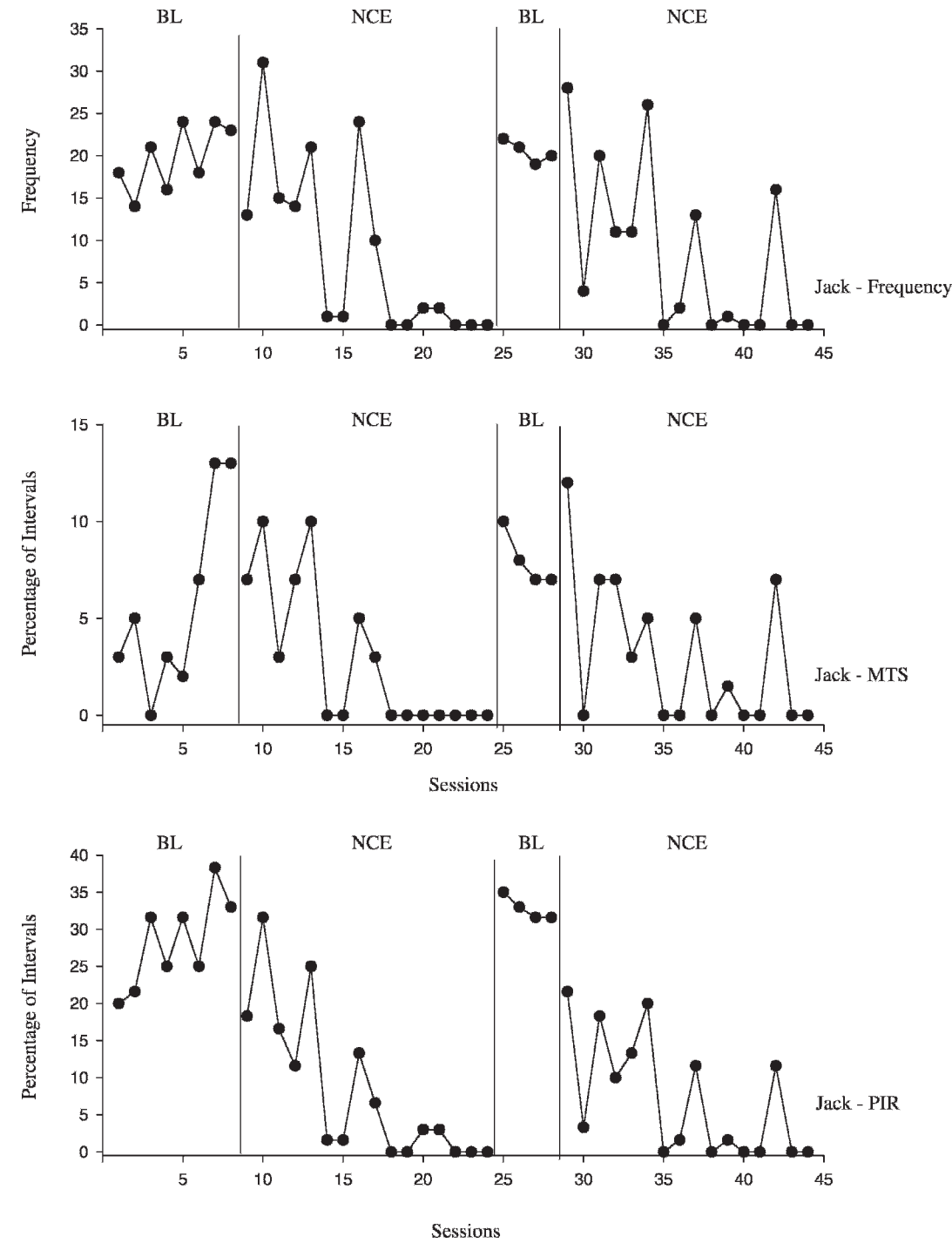


Figure 5. Intervention data (NCE with extinction) based on duration, MTS, and PIR for Jack.

the baseline mean line in the expected direction. For the purpose of behavior reduction, those data points would have to fall below the two criterion lines.

After the structured criterion method was applied to each of the 15 graphs, data were analyzed by comparing whether a treatment effect was identified for each of the data sets. If a treatment effect or no treatment effect was scored for both a discontinuous (MTS or PIR) and a continuous (duration or frequency) method, then an agreement was indicated. If a treatment effect was scored for one method (discontinuous or continuous) but not for the other method (discontinuous or continuous), then a disagreement was indicated. The first author manually applied the structured criterion method to each of the treatment assessment graphs to formulate a decision on whether a functional relation was demonstrated. Interobserver agreement data were collected by having the second author independently apply the structured criterion method to the same set of figures to determine whether a functional relation was demonstrated. Agreement was calculated by dividing the total number of agreements by the number of agreements plus disagreements and multiplying by 100%. Agreement data were collected for 53% of the samples and averaged 100%.

## RESULTS

Results of the expert panel interview and the structured criterion method are depicted in Figure 6. The top panel shows the percentage agreement when MTS and PIR were compared to duration. During the expert panel interview, the number of agreements for respondents' answers regarding the presence or absence of functional control was high for both MTS and PIR when compared to duration. However, slightly more agreements were obtained for MTS (24) than for PIR (21). These findings indicate that similar treatment interpretations regarding functional control during a treatment

assessment were made when using both MTS and PIR for estimating duration; however, similar outcomes were obtained more often when using MTS than when using PIR. During the structured criterion method, the number of agreements for whether a functional relation was obtained was the same for both MTS (two) and for PIR (two).

The bottom panel of Figure 6 shows the percentage agreement when MTS and PIR were compared to frequency. During the expert panel interview, the number of agreements for respondents' answers regarding the presence or absence of functional control was high for both MTS and PIR when compared to frequency. However, slightly more agreements were obtained for PIR (16) than for MTS (15). During the structured criterion method, the number of agreements for whether a functional relation was obtained was higher for PIR (two) than for frequency (one). These findings indicate that treatment interpretations similar to those made when analyzing frequency data regarding functional control during a treatment assessment were made when using both MTS and PIR; however, similar outcomes were obtained slightly more often when using PIR than when using MTS.

## DISCUSSION

In the current study, we extended previous research by comparing MTS and PIR to continuous duration recording of stereotypy across treatment assessment conditions, to determine whether the recording method used might affect data interpretation. Results showed that when comparing treatment interpretations from data measured using duration to those measured using MTS and PIR, those using MTS better corresponded with duration than those using PIR based on the expert interview. However, no difference was observed between MTS and PIR when using the structured criterion method.

We also compared MTS and PIR to frequency recording of self-injurious behavior

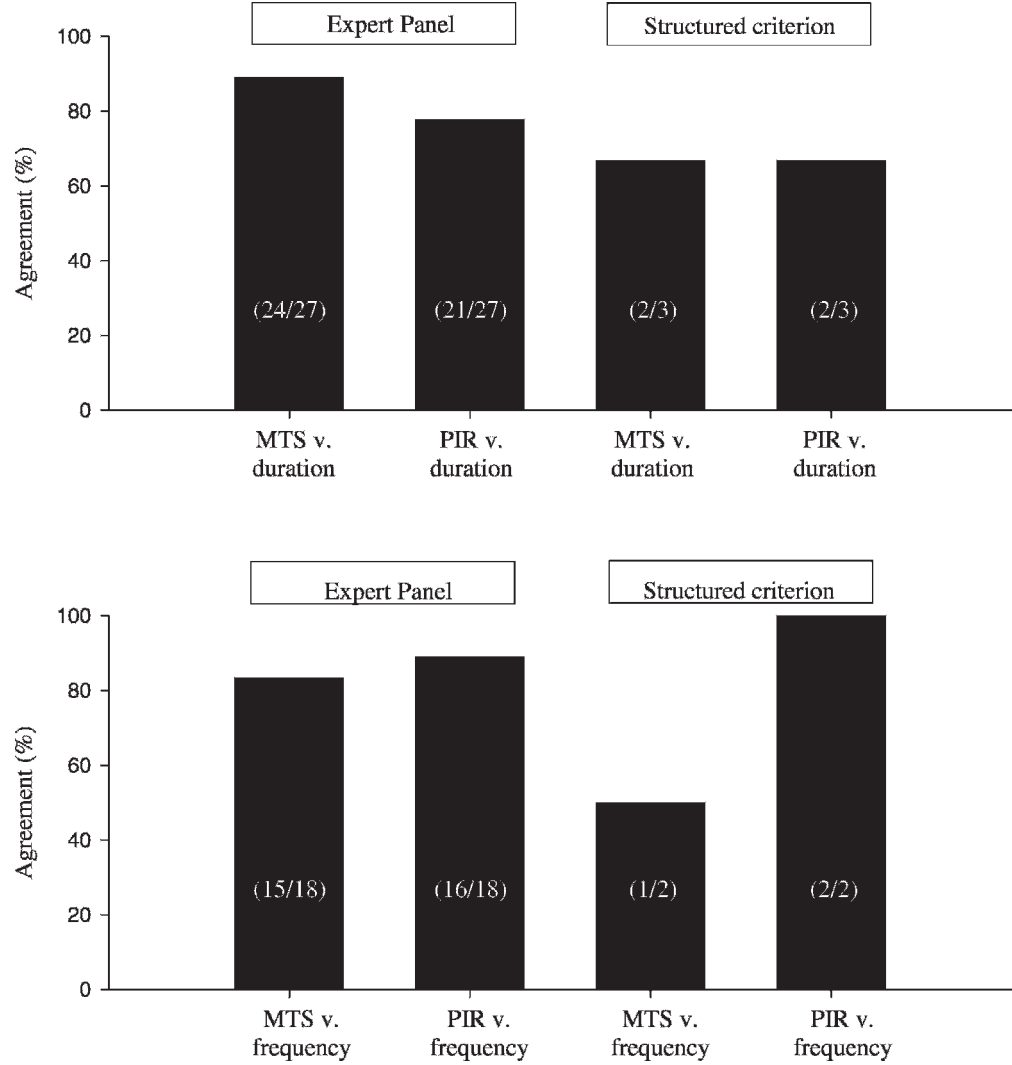


Figure 6. Correspondence data from the expert panel interview and the structured criterion method when comparing MTS and PIR to duration and to frequency.

across treatment analysis conditions to determine whether the recording method used would affect treatment interpretation. When comparing treatment interpretations from data sets depicting MTS and PIR to those depicting frequency, data depicting PIR yielded a slightly greater percentage agreement than those depicting MTS during the expert panel, and data depicting PIR yielded a much larger percentage agreement than those depicting MTS during the structured criterion method. However, it is

important to note that this involved only two comparisons in which PIR agreed with frequency both times and MTS agreed with frequency once. Thus, it is unclear whether similar findings would have been obtained if more data sets had been included. Closer examination of the treatment interpretation data revealed the types of errors that were made across both the expert interview and structured criterion methods. When using the expert interview, 25% of the errors made

with PIR were false positives (i.e., a functional relation was detected when one was not present) and 75% of the errors made with PIR were false negatives (i.e., a functional relation was not detected when one was present). For errors made with MTS, 33% were false positives and 67% were false negatives. When using the structured criterion method, 100% of the errors made with both MTS and PIR were false negatives. The present findings indicate that both MTS and PIR were more likely to result in false-negative interpretations than false-positive interpretations for identifying functional relations. This aspect of the data suggests that errors made in identifying functional relations due to the use of MTS or PIR are not random. When using MTS or PIR, one may be less likely to detect functional relations that are evident when continuous observation methods are used, leading to the discontinuation of a potentially effective treatment. This finding suggests that discontinuous measurement systems may be contraindicated for the detection of small treatment effects because, given the likelihood of a false-negative interpretation, a functional relation may remain undetected.

The data indicated that for the three errors obtained when comparing MTS and duration, one was a false positive and two were false negatives. For the six errors obtained when comparing PIR and duration, all were false negatives. For the three errors obtained when comparing MTS and frequency, one was a false positive and two were false negatives. For the two errors obtained when comparing PIR and frequency, both were false positives. Given these findings, researchers and clinicians should be aware of a likelihood of false-positive findings when using PIR to estimate frequency and the possibility of false-negative findings when using PIR to estimate duration. MTS generally produced false-negative results for both duration and frequency.

Previous research has shown that discontinuous measurement methods may result in over-

or underestimations of the continuous data record to varying degrees. However, it is unclear how differences in correspondence may result in different conclusions about functional relations. The present study illustrated a useful method for evaluating measurement procedures (i.e., by evaluating differences in treatment interpretations regarding functional control during an ABAB treatment analysis). Information obtained through this method provides a better indicator of how differences in measurement may alter conclusions made when designing treatment programs for behavior reduction.

In addition, the present study illustrated the utility of the structured criterion method (as proposed by Fisher et al., 2003) for detecting functional relations across a variety of target responses and treatment procedures. Results showed that the structured criterion method yielded similar outcomes to those obtained by the expert review panel. These findings demonstrate the generality of the structured criterion method for aiding in visual inspection of treatment effects.

Based on these findings, the recommended discontinuous measurement method for estimating target responses appropriately measured using duration is MTS, and the recommended discontinuous measurement method for estimating the frequency of responses is PIR. However, these findings should be interpreted with caution because the small number of participants included may have limited their generality. In the present study, the interval size used was 10 s, and a variety of different response frequencies and bout durations were observed across participants and experimental conditions. For Amy, an average of nine responses with a mean bout duration of 30 s were observed during baseline sessions and an average of 66 responses with a mean bout duration of 2 s were observed during treatment sessions. For Daniel, an average of nine responses with a mean bout duration of 4 s was observed during baseline sessions and an

average of five responses with a mean bout duration of 2 s was observed during treatment sessions. For Beth, an average of three responses with a mean bout duration of 106 s was observed during baseline sessions and an average of 24 responses with a mean bout duration of 1 s was observed during treatment sessions. For Jack's NCR treatment assessment, an average of 48 responses with a mean bout duration of 1 s was observed during baseline sessions, and an average of six responses with a mean bout duration of 1 s was observed during treatment sessions. For Jack's NCE treatment assessment, an average of 20 responses with a mean bout duration of 1 s was observed during baseline sessions, and an average of eight responses with a mean bout duration of 1 s was observed during treatment sessions. Future research might compare outcomes obtained by MTS and PIR for detecting functional relations during a treatment assessment across a larger number of data sets that incorporate a variety of different treatment assessments and response topographies. Because different response topographies are characterized by different frequencies and bout durations, the current study could be extended by specifically selecting response topographies characterized by low, moderate, and high response frequencies and short, medium, and long bout durations to determine which discontinuous measurement method is most appropriate for a given response frequency or duration.

Also, in the current study we evaluated whether similar outcomes would be obtained when evaluating treatment for reducing problem behavior. It is possible that different

findings would be obtained when examining treatment procedures aimed at increasing behavior. Future research might address this issue by comparing outcomes obtained using MTS and PIR to those obtained when using continuous recording for detecting functional relations in the context of a skill-acquisition program (e.g., shaping or chaining).

## REFERENCES

- Ahearn, W. H., Clark, K. M., MacDonald, R. P. F., & Chung, B. I. (2007). Assessing and treating vocal stereotypy in children with autism. *Journal of Applied Behavior Analysis*, 40, 263–275.
- Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis*, 36, 387–406.
- Gardenier, N. C., MacDonald, R., & Green, G. (2004). Comparison of direct observational methods for measuring stereotypic behavior in children with autism spectrum disorders. *Research in Developmental Disabilities*, 25, 99–118.
- Harrop, A., & Daniels, M. (1986). Methods of time sampling: A reappraisal of momentary time sampling and partial interval recording. *Journal of Applied Behavior Analysis*, 19, 73–76.
- Murphy, G., & Goodall, E. (1980). Measurement error in direct observations: A comparison of common recording methods. *Behaviour Research and Therapy*, 18, 147–150.
- Powell, J., Martindale, A., & Kulp, S. (1975). An evaluation of time-sample measures of behavior. *Journal of Applied Behavior Analysis*, 8, 463–469.
- Suen, H. K., Ary, D., & Covalt, W. (1991). Reappraisal of momentary time sampling and partial-interval recording. *Journal of Applied Behavior Analysis*, 24, 803–804.

*Received July 12, 2006*

*Final acceptance December 18, 2006*

*Action Editor, James Carr*